WHAT IS CLAIMED IS:

1. A method for generating training data that can be used with statistical models to normalize abbreviations in text, including:

providing a corpus of text including expansions of the abbreviations to be normalized;

identifying the expansions in the corpus of text;

generating context information describing the context of the text in which the expansions were identified; and

storing training data as a function of the context information.

2. The method of claim 1 wherein:

generating context information includes generating local context information; and

storing training data includes storing local context data as training data.

3. The method of claim 2 wherein the local context information and local context training data includes sentence level information.

4. The method of claim 3 wherein the sentence level information includes words in a sentence in which the identified expansion is located.

5. The method of claim 1 wherein:

generating context information includes generating discourse context information; and

storing training data includes storing discourse context data as training data.

6. The method of claim 5 wherein the discourse context information and discourse context training data include text section level information.

7. The method of claim 1 wherein:

generating context information includes generating local context information and discourse context information; and

storing training data includes storing local context data and discourse data as training data.

8. The method of claim 1 wherein storing the training data includes storing a set of feature vectors, each feature vector including the context information generated for the associated expansion identified in the corpus of text.

9. The method of claim 8 and further including processing text using a Maximum Entropy model and the stored feature vectors to normalize abbreviations in the text.

10. The method of claim 8 wherein each feature vector further includes the abbreviation and associated expansion.

11. The method of claim 1 and further including processing text using a statistical model and the stored training data to normalize abbreviations in the text.

12. The method of claim 1 wherein:

the method further includes providing stored abbreviation data representative of abbreviations and associated expansions for which training data is to be generated; and

identifying the expansions includes processing the corpus of text as a function of the stored abbreviation data.

13. A method for electronically generating feature vectors that can be used in connection with electronic data processing systems implementing statistical models to normalize abbreviations in text, including:

providing a database of abbreviation data representative of abbreviations and associated expansions to be normalized;

providing a database having a corpus of text including expansions of the abbreviations to be normalized;

processing the corpus of text as a function of the abbreviation data to identify the expansions in the corpus of text;

generating context information describing the context of the text in which the expansions were identified; and

storing a set of feature vectors, each feature vector including the context information generated for the associated expansion identified in the corpus of text.

14. The method of claim 13 wherein the feature vectors include local level context information and discourse level context information.

15. The method of claim 13 and further including operating an electronic data processing system implementing a statistical model and the stored set of feature vectors to normalize abbreviations in the text.